ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Applied Logic 3 (2005) 308–328

JOURNAL OF APPLIED LOGIC

www.elsevier.com/locate/jal

A comprehensive semantic framework for data integration systems

Andrea Calì [a,*], Domenico Lembo [b], Riccardo Rosati [b]

[a] *Faculty of Computer Science, Free University of Bolzano/Bozen, Italy*
[b] *Dipartimento di Informatica e Sistemistica, Università di Roma "La Sapienza", Italy*

Available online 21 August 2004

Abstract

A data integration system provides the user with a unified view, called *global schema*, of the data residing at different *sources*. Users issue their queries against the global schema, and the system computes answers to queries by suitably accessing the sources, through the *mapping*, i.e., the specification of the relationship between the global schema and the sources. Since sources are in general autonomous subsystems, the information provided by the data at the sources and the mapping is likely not to be consistent with the knowledge expressed by the global schema. Therefore, the question arises of how to interpret user queries in such a situation, i.e., in the presence of data contradicting the global schema and the mapping. In this paper, we provide an in-depth analysis of the problem of dealing with inconsistencies in data integration systems. In this respect, we highlight the central role played by the mapping, and propose a general "mapping-centered" semantics that allows for computing significant answers to user queries even in the presence of inconsistent information. Based on such a semantic analysis, we define a general formal framework for data integration. Then, we argue that our semantic approach formalizes a very reasonable way of handling inconsistency in such systems, since practically all the existing proposals in the literature can be reconstructed in our framework. This allows for comparing and evaluating the different existing proposals.

Keywords: Data integration; Data inconsistency; Repair semantics; Consistent query answering

* Corresponding author.
E-mail addresses: cali@inf.unibz.it (A. Calì), lembo@dis.uniroma1.it (D. Lembo), rosati@dis.uniroma1.it (R. Rosati).

1570-8683/$ – see front matter
doi:10.1016/j.jal.2004.07.022

1. Introduction

The task of a data integration system is to combine the data residing at different sources, and providing the user with a unified view of these data, called *global schema* [32,34, 38,44]. Users query the global schema, while the system carries out the task of suitably accessing different sources and assembling the data retrieved at each source into the final answer to the query.

The global schema is therefore the interface by which users issue their queries to the system. The system answers the queries by accessing the appropriate sources, thus freeing the user from the knowledge on where data are, and how data are structured at the sources. Notably, sources are in general autonomous systems that can be accessed through different modalities.

The interest in this kind of systems has been continuously growing in the last years. Many organizations face the problem of integrating data residing in several sources. Companies that build a Data Warehouse, a Data Mining, or an Enterprise Resource Planning system must address this problem. Also, integrating data in the World Wide Web is the subject of several investigations and projects nowadays. Finally, applications requiring accessing or re-engineering legacy systems must deal with the problem of integrating data stored in pre-existing sources.

A central aspect of a data integration system is the specification of the relationship between the global schema and the sources; such a specification is given in the form of a so-called *mapping*. Two kinds of mapping are commonly adopted in the literature: the *global-as-view* mapping, in which every element of the global schema is associated with a view over the sources, and the *local-as-view* mapping, which requires the sources to be defined as views over the global schema [38,40,44].

Summarizing, the high-level structure of a data integration system that is commonly adopted consists of a triple $\langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$, where $\mathcal{G}$ is the global schema, $\mathcal{S}$ is the set of sources and $\mathcal{M}$ is the mapping. All such components correspond to logical theories. Therefore, the meaning of a data integration system is given through the semantics of the logical theory corresponding to the system specification. Since all current approaches to data integration use (fragments of) first-order logic to specify the global schema, the mapping, and the sources, the semantics of a data integration system is in general defined in terms of the classical, first-order semantics of a first-order theory.

However, such an approach to the semantics of data integration system is not satisfactory. Indeed, as already mentioned, sources are in general autonomous subsystems, hence the information provided by the data at the sources and the mapping are likely not to be consistent with the knowledge expressed by the global schema [13,26]. In these cases, the first-order semantics of the system simply states that there is no model for the system: such an "empty" meaning is not appropriate, since one would like to

(i) be able to derive significant information from a data integration system even in the presence of inconsistency, a capability that is not provided by such semantics;
(ii) treat different forms of inconsistency in different ways, while the first-order semantics gives the same, empty meaning to all kinds of inconsistency.

These issues are well-known limitations of classical logic that have been studied in the literature in paraconsistent logics, belief revision and nonmonotonic reasoning [19,23].

In order to overcome this semantic problem, we have to answer the following crucial question: *what is the meaning of a data integration system in the presence of inconsistency?* Since the main task of a data integration system is to provide answers to user queries, such a question can also be formulated as follows: what are the answers to be returned to a user query in the presence of inconsistency?

This issue has been recently addressed in the field of *inconsistent databases*: in this setting, the central problem is computing "consistent" answers to queries posed to databases in which data do not satisfy the database schema, which contains a set of integrity constraints [4,12,29,41].

All approaches in this setting are based on the following principle: *schema is stronger than data*. In other words, the database schema (i.e., the set of integrity constraints) is considered as the actually reliable information (strong knowledge), while data are considered as information to be revised (weak knowledge). Therefore, the problem amounts to deciding how to "repair" (i.e., change) data in order to reconcile them with the information expressed in the schema.

Notably, the above principle is an even more natural assumption in data integration, where, due to the autonomous nature of the sources, data may not be completely reliable and/or reconciled, while the global schema provides a reliable specification of the semantics of data.

Even though the essence of the semantic problem that arises in inconsistent databases is the same as the one illustrated for data integration, the different structure of a data integration system with respect to a single database (in particular, the presence of autonomous data sources and of the mapping) makes the problem in the data integration setting significantly harder to deal with. However, the first attempts to define a semantics for data integration systems in the presence of inconsistency in general have tried to extend, in a more or less "intuitive" way, semantic approaches that had been previously defined for inconsistent databases.

In this paper we try to provide a rigorous study of the problem of dealing with inconsistency in data integration systems. We address the problem in a very general and comprehensive setting, that amplifies the structural differences with the single database setting. Indeed, we want to be able to deal with very expressive global schema specifications and mapping assertions: therefore, we use first-order logic to represent such components of a data integration system.

More specifically:

- we consider the well-established logic-based formalization of data integration systems (see e.g. [38]), and restate it in terms of first-order logic. Such a framework is very general, since it is able to capture the main logical approaches to data integration proposed so far. Among other things, such a generality allows us to compare and evaluate the different existing proposals;
- we provide an in-depth analysis of the problem of dealing with inconsistencies in data integration systems. In this respect, we highlight the central role played by the mapping, and propose a general "mapping-centered" semantics that allows for computing

significant answers to user queries even in the presence of inconsistent information. We argue that our semantic approach formalizes a very reasonable way of handling inconsistency in such systems, since all the existing proposals in the literature can be reconstructed in our semantic framework.

The paper is structured as follows. In Section 2, we provide the syntax and the first-order semantics of the formal framework for data integration. In Section 3, we study the problem of dealing with inconsistency in data integration systems, and provide new formal semantics for the integration framework. In Section 4, we analyze the state of the art in inconsistent databases and data integration, and show that our framework is able to capture all the main approaches to consistent query answering in database and data integration systems proposed so far. Finally, we conclude the paper in Section 5.

2. Framework

In this section we define a general formal framework for data integration. Informally, a data integration system consists of a (virtual) global schema, which specifies the global elements exposed to the user, a source schema, which describes the structure of the sources in the system, and a mapping, which specifies the relationship between the sources and the global schema. User queries are posed on the global schema, and the system provides the answers to such queries by exploiting the information supplied by the mapping and accessing the sources that contain relevant data. Thus, from the syntactic viewpoint, the specification of an integration system depends on the following parameters:

- The form of the global schema, i.e., the formalism used for expressing global elements and relationships between global elements, e.g., integrity constraints expressed over a database schema. Several settings have been considered in the literature, where, for instance, the global schema can be relational [28], object-oriented [8], semi-structured [42], based on Description Logics [16,36], etc.
- The form of the source schema, i.e., the formalism used for expressing data at the sources and relationships between such data. In principle, the formalisms commonly adopted for the source schema are the same as those mentioned for the global schema;
- The form of the mapping. Two basic approaches have been proposed in the literature, called respectively *global-as-view* (GAV) and *local-as-view* (LAV) [40,44]. The GAV approach requires that the global schema is defined in terms of the data sources: more precisely, every element of the global schema is associated with a view, i.e., a query, over the sources, so that its meaning is specified in terms of the data residing at the sources. Conversely, in the LAV approach, the meaning of the sources is specified in terms of the elements of the global schema: more exactly, the mapping between the sources and the global schema is provided in terms of a set of views over the global schema, one for each source element.
- The language of the mapping, i.e., the query language used to express views in the mapping.

– The language of the user queries, i.e., the query language adopted by users to issue queries on the global schema.

Let us now turn our attention to the semantics. According to [38], the semantics of a data integration system is given in terms of the extension of the elements of the global schema (e.g., one set of tuples for each global relation if the global schema is relational, one set of objects for each global class if it is object-oriented, etc.). Such extension has to satisfy (i) the knowledge expressed by the global schema, and (ii) the mapping specified between the global and the source schema.

Roughly speaking, the notion of satisfying the mapping depends on how the data retrieved from the sources are interpreted with respect to the data that satisfy the global schema. Different interpretations lead to different notions. More specifically, when the mapping is GAV, data that satisfy each global element can be considered a superset or a subset of the data retrieved by the associated view over the sources. In the case of LAV mapping, data stored in each source element can be considered a subset or a superset of the data that satisfy the corresponding view over the global schema. Both in GAV and in LAV, views in the mapping are called *sound* in the former case and *complete* in the latter. A view can be also considered sound and complete at the same time: in this case it is called *exact*. When all views are sound (resp. complete, exact), the mapping is called sound (resp. complete, exact).

In the following, we provide a precise characterization of the concepts informally explained above. In particular, we define a logical formal framework which captures all the syntactic and semantic aspects of data integration applications. In our framework, the languages used to specify the global and the source schema, the mapping and user queries rely on first-order logic (FOL). Actually, the expressive power of FOL allows us to capture most of the approaches to data integration proposed in the literature. Moreover, in the spirit of [38], we consider mappings of a very general form, which allows for specifying GAV and LAV mappings as special cases. For clearness of presentation, we first address the syntax and then the semantics of our framework.

2.1. Syntax

A data integration system $\mathcal{I}$ is a triple $\langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$, where:

– $\mathcal{G}$ is the *global schema*, expressed in some subset of FOL with equality on the alphabet formed by a possibly infinite set Γ of constant symbols, and a set $\mathcal{A}_\mathcal{G}$ of predicate (or relation) symbols with associated arity (we do not consider functions in this paper). In other words, $\mathcal{G}$ is composed by a set of predicates and a set of first-order sentences on such predicates.
– $\mathcal{S}$ is the *source schema*, composed by the schemas of the various sources. We assume that the source schema is simply a set of predicate (or relation) symbols with associated arity of an alphabet $\mathcal{A}_\mathcal{S}$. In other words, we do not allow for the specification of FOL sentences establishing integrity constraints over data sources. This implies that data stored at the sources are always considered locally consistent. This is a common as-

sumption in data integration, because sources are in general autonomous and external to the integration system, which is not in charge to analyze their consistency.

– $\mathcal{M}$ is the *mapping* between $\mathcal{G}$ and $\mathcal{S}$. It is constituted by a set of *assertions* in which, intuitively, views, i.e., queries, expressed over $\mathcal{G}$ are put in correspondence to queries expressed over $\mathcal{S}$. We assume that queries in the mapping are FOL queries, i.e., open formulas of the form

$$\{x_1, \ldots, x_n \mid \phi(x_1, \ldots, x_n)\}, \tag{1}$$

where $x_1, \ldots, x_n$ is the sequence of free variables of ϕ, and n is the *arity* of the query. More precisely, a mapping assertion assumes one of the following forms

$$q_{\mathcal{S}} \sqsubseteq q_{\mathcal{G}},$$
$$q_{\mathcal{G}} \sqsubseteq q_{\mathcal{S}},$$

where $q_{\mathcal{S}}$ and $q_{\mathcal{G}}$ are two queries of the same arity, respectively over the alphabet $\mathcal{A}_{\mathcal{S}} \cup \Gamma$ and the alphabet $\mathcal{A}_{\mathcal{G}} \cup \Gamma$.
We point out that the above definition corresponds to a generalized form of mapping that comprises LAV and GAV as special cases. Indeed, the GAV approach corresponds to restricting the queries $q_{\mathcal{G}}$ to single atom queries, i.e., queries containing a single element of the global schema, whereas the LAV approach corresponds to restricting the queries $q_{\mathcal{S}}$ to queries containing a single element of the source schema.

Finally, we consider *user queries* posed to a data integration system $\mathcal{I}$, and define their syntax. Each such query q is a formula that is intended to provide the specification of which data to extract from the integration system. We assume that user queries are first-order queries, i.e., formulas of form (1), over the alphabet $\mathcal{A}_{\mathcal{G}} \cup \Gamma$.

Example 1. Consider a data integration system $\mathcal{I}_0 = \langle \mathcal{G}_0, \mathcal{S}_0, \mathcal{M}_0 \rangle$, where the global schema alphabet $\mathcal{A}_{\mathcal{G}_0}$ comprises the three binary relation symbols *DeptDirector*, *EmployeeDept* and *DeptLocation*, which respectively indicate director of departments, department of employees, and location of departments. Assume that the following FOL sentences are specified over the alphabet $\mathcal{A}_{\mathcal{G}_0}$,

$$\forall x, y_1, y_2 . \mathit{EmployeeDept}(x, y_1) \wedge \mathit{EmployeeDept}(x, y_2) \supset y_1 = y_2,$$
$$\forall x, y . \mathit{DeptDirector}(x, y) \supset \mathit{EmployeeDept}(y, x),$$

which state respectively that an employee works in only one department, and that a director of a department is also an employee of the same department.

Consider now the source schema $\mathcal{S}_0$, and assume that its alphabet $\mathcal{A}_{\mathcal{S}_0}$ comprises the three binary relation symbols *IsBossOf*, *IsMemberOf* and *WorksIn*, which respectively specify bosses of employees, members of departments, and cities in which employees work.

According to the above description of the sources, we define the mapping $\mathcal{M}_0$ with the following three assertions:

$$\{x, y \mid \mathit{DeptDirector}(x, y)\} \sqsubseteq \{x, y \mid \exists z . \mathit{IsBossOf}(y, z) \wedge \mathit{IsMemberOf}(z, x)\},$$

$$\{x, y \mid \exists z.\mathit{IsBossOf}(y, z) \wedge \mathit{IsMemberOf}(z, x)\} \sqsubseteq \{x, y \mid \mathit{DeptDirector}(x, y)\},$$
$$\{x, y, z \mid \mathit{IsMemberOf}(x, y) \wedge \mathit{WorksIn}(x, z)\} \sqsubseteq \{x, y, z \mid \mathit{EmployeeDept}(x, y) \wedge \mathit{DeptLocation}(y, z)\}.$$

Finally, consider the following query issued on the global schema

$$\{x, y \mid \mathit{EmployeeDept}(x, y)\},$$

which asks for the pairs employee-department.

2.2. Semantics

For the sake of simplicity of presentation, we assume that the domain of interpretation is a fixed denumerable set of elements Δ and that every such element is denoted uniquely by a constant symbol, called its *standard name* [39]. We assume that the set of standard names is the set of constants Γ previously introduced. Therefore, without loss of generality we assume that $\Delta = \Gamma$. We point out that, in our framework, we can also adopt the finite model assumption, i.e., we can assume that Δ is a finite set. Actually, the study of both finite and unrestricted models is relevant in database theory.

Intuitively, to specify the semantics of a data integration system, we have to start with a set of data at the sources, and we have to specify which are the data that satisfy the global schema with respect to such data at the sources. Thus, in order to assign the semantics to a data integration system $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$, we start by considering a *source model* for $\mathcal{I}$, i.e., an interpretation $\mathcal{D}$ for the source schema $\mathcal{S}$. Moreover, we assume that each instance of the information sources to be integrated has only one model. This is a classical assumption in data integration, since the information sources to be integrated are typically databases, i.e., they provide the integration system with a single fixed database extension. Therefore, in the following, with a little abuse of notation, we use the symbol $\mathcal{D}$ to denote both the source instance and the unique model of such an instance.

Based on $\mathcal{D}$, we now specify which is the information content of the global schema $\mathcal{G}$. We call any interpretation over Δ of the symbols in $\mathcal{A}_\mathcal{G}$ a *global interpretation* for $\mathcal{I}$.

Definition 1. Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a data integration system, let $\mathcal{D}$ be a source model for $\mathcal{I}$, a global interpretation $\mathcal{W}$ for $\mathcal{I}$ is a *model for $\mathcal{I}$ with respect to $\mathcal{D}$* iff

(1) $\mathcal{W}$ is a model of $\mathcal{G}$, i.e., $\mathcal{W} \models \mathcal{G}$;
(2) $\mathcal{W}$ satisfies the mapping $\mathcal{M}$ with respect to $\mathcal{D}$. More precisely, we say that $\mathcal{W}$ *satisfies $\mathcal{M}$ with respect to $\mathcal{D}$* if:
   – for each assertion in $\mathcal{M}$ of the form $q_\mathcal{S} \sqsubseteq q_\mathcal{G}$,

$$q_\mathcal{S}^\mathcal{D} \subseteq q_\mathcal{G}^\mathcal{W},$$

where $q_\mathcal{S}^\mathcal{D}$ (resp., $q_\mathcal{G}^\mathcal{W}$) denotes the result of evaluating $q_\mathcal{S}$ (resp., $q_\mathcal{G}$) over the interpretation $\mathcal{D}$ (resp., $\mathcal{W}$), i.e., the set of tuples of elements of Δ associated to the free variables of $q_\mathcal{S}$ (resp., $q_\mathcal{G}$) by the interpretation $\mathcal{D}$ (resp., $\mathcal{W}$). In other words, an assertion of the form $q_\mathcal{S} \sqsubseteq q_\mathcal{G}$ is satisfied if each tuple in $q_\mathcal{S}^\mathcal{D}$ is also a tuple of $q_\mathcal{G}^\mathcal{W}$;

– for each assertion in $\mathcal{M}$ of the form $q_{\mathcal{G}} \sqsubseteq q_{\mathcal{S}}$,

$$q_{\mathcal{G}}^{\mathcal{W}} \subseteq q_{\mathcal{S}}^{\mathcal{D}},$$

i.e., each tuple in $q_{\mathcal{G}}^{\mathcal{W}}$ is also a tuple of $q_{\mathcal{S}}^{\mathcal{D}}$.

The set of all models for $\mathcal{I}$ with respect to $\mathcal{D}$ is called *the semantics of $\mathcal{I}$ with respect to $\mathcal{D}$*, denoted by $sem(\mathcal{I}, \mathcal{D})$.

Notice that, from the above semantics of the mapping $\mathcal{M}$, it follows that in our framework it is possible to express the sound, the complete, and the exact interpretation of the mapping assertions studied in data integration [38]. In particular, if we want to formulate a generic mapping assertion A defining a relationship between the query $q_{\mathcal{G}}$ over the global schema and the query $q_{\mathcal{S}}$ over the source schema:

– a *sound* interpretation of A corresponds in our framework to the assertion $q_{\mathcal{S}} \sqsubseteq q_{\mathcal{G}}$;
– a *complete* interpretation of A corresponds to the assertion $q_{\mathcal{G}} \sqsubseteq q_{\mathcal{S}}$;
– an *exact* interpretation of A corresponds to the pair of assertions $q_{\mathcal{S}} \sqsubseteq q_{\mathcal{G}}$, $q_{\mathcal{G}} \sqsubseteq q_{\mathcal{S}}$.

Let us now turn our attention to queries. In order to define the semantics of a query q over a data integration system $\mathcal{I}$, we have to take into account all the models of $\mathcal{I}$ with respect to $\mathcal{D}$.

Definition 2. Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a data integration system, let $\mathcal{D}$ be a source model for $\mathcal{I}$, and let q be a user query over $\mathcal{I}$, then the set of *certain answers* of q with respect to $\mathcal{I}$ and $\mathcal{D}$, denoted by $ans(q, \mathcal{I}, \mathcal{D})$, is defined as follows:

$$ans(q, \mathcal{I}, \mathcal{D}) = \left\{ \langle c_1, \ldots, c_n \rangle \mid \text{for each } \mathcal{W} \in sem(\mathcal{I}, \mathcal{D}), \langle c_1, \ldots, c_n \rangle \in q^{\mathcal{W}} \right\}.$$

Such a notion of answers, corresponding to skeptical entailment, is the most used in data integration; however the notion of *possible answers*, corresponding to credulous entailment, can also be defined [32,38].

Example 1 (*contd.*). Assume now that the set of constants Γ contains, among others, the elements John, Mary, D1, NewYork, and consider the following source model $\mathcal{D}_0$ for $\mathcal{I}_0$,

$$\mathcal{D}_0 = \big\{ IsBossOf(\mathsf{John}, \mathsf{Mary}), IsMemberOf(\mathsf{Mary}, \mathsf{D1}), \\ WorksIn(\mathsf{John}, \mathsf{NewYork}), WorksIn(\mathsf{Mary}, \mathsf{New\ York}) \big\}.$$

Then, in each global interpretation that satisfies $\mathcal{M}_0$ with respect to $\mathcal{D}_0$ the following set $\mathcal{W}_0$ of facts holds,

$$\mathcal{W}_0 = \big\{ DeptDirector(\mathsf{D1}, \mathsf{John}), EmployeeDept(\mathsf{Mary}, \mathsf{D1}), \\ DeptLocation(\mathsf{D1}, \mathsf{New\ York}) \big\}.$$

The set $\mathcal{W}_0$ and the global sentence $\forall x, y.DeptDirector(x, y) \supset EmployeeDept(y, x)$ entail the fact $EmployeeDept(\mathsf{John}, \mathsf{D1})$ (i.e., if John is the director of department D1,

then John is also an employee of D1). Furthermore, this fact can be added to $\mathcal{W}_0$ without affecting the satisfaction of the mapping. Therefore,

$$sem(\mathcal{I}_0, \mathcal{D}_0) = \{\mathcal{W} \mid \mathcal{W} \models \mathcal{G}_0 \text{ and } \mathcal{W} \supseteq \mathcal{W}_0 \cup \{EmployeeDept(\mathsf{John}, \mathsf{D1})\}\}.$$

Then, for the query $q = \{x, y \mid EmployeeDept(x, y)\}$, we have that

$$ans(q, \mathcal{I}_0, \mathcal{D}_0) = \{\langle \mathsf{Mary}, \mathsf{D1}\rangle, \langle \mathsf{John}, \mathsf{D1}\rangle\}.$$

3. General semantics

According to the semantics $sem(\mathcal{I}, \mathcal{D})$, it may be the case that the data retrieved from the sources cannot be reconciled in the global schema in such a way that both the knowledge in the global schema and the mapping are satisfied [37]. In such cases, $sem(\mathcal{I}, \mathcal{D}) = \emptyset$, therefore, by Definition 2, every tuple is in the answer set of every query. This is not an acceptable way of handling inconsistency: as motivated by the studies on consistent query answering in inconsistent databases [4,12,29], it could be possible to derive significant answers to queries even in the presence of inconsistency.

Example 2 (*contd.*). Consider now the following source model $\mathcal{D}'_0$ for $\mathcal{I}_0$,

$$\begin{aligned}\mathcal{D}'_0 = \{&IsBossOf(\mathsf{John}, \mathsf{Mary}), IsMemberOf(\mathsf{Mary}, \mathsf{D1}),\\ &IsMemberOf(\mathsf{John}, \mathsf{D2}), WorksIn(\mathsf{John}, \mathsf{NewYork}),\\ &WorksIn(\mathsf{Mary}, \mathsf{New\ York})\},\end{aligned}$$

where D2 is a new symbol of Γ.

Proceeding as before, we have now that, in each global interpretation that satisfies $\mathcal{M}_0$ with respect to $\mathcal{D}'_0$, the following set $\mathcal{W}'_0$ of facts holds,

$$\begin{aligned}\mathcal{W}'_0 = \{&DeptDirector(\mathsf{D1}, \mathsf{John}), EmployeeDept(\mathsf{Mary}, \mathsf{D1}),\\ &DeptLocation(\mathsf{D1}, \mathsf{New\ York}), EmployeeDept(\mathsf{John}, \mathsf{D2}),\\ &DeptLocation(\mathsf{D2}, \mathsf{New\ York})\}.\end{aligned}$$

Furthermore, analogously to the previous case, $\mathcal{W}'_0$ and the global sentence $\forall x, y.$ $DeptDirector(x, y) \supset EmployeeDept(y, x)$ entail the fact $EmployeeDept(\mathsf{John}, \mathsf{D1})$. Such fact, together with $EmployeeDept(\mathsf{John}, \mathsf{D2})$, which is contained in $\mathcal{W}'_0$, violates the sentence of $\mathcal{G}_0$ stating that an employee works in only one department. On the other hand, the mapping $\mathcal{M}_0$ and the other sentence in $\mathcal{G}_0$ force us to consider in the semantics of the system those global interpretations of $\mathcal{G}_0$ in which both such facts hold. Therefore, $sem(\mathcal{I}_0, \mathcal{D}'_0) = \emptyset$, i.e., the system $\mathcal{I}_0$ is inconsistent with respect to $\mathcal{D}'_0$, and the certain answers to each query of arity n are all the n-tuples of elements of Γ.

Roughly speaking, query answering under the classical *sem* is not significant in the presence of inconsistency, since the system provides answers to user queries which are returned only because of the "ex falso quodlibet" principle, but which are not "positively" supported by data stored at the sources. In our scenario, for example, all pairs of elements of Γ are in

the answer set of the query $\{x, y \mid \mathit{EmployeeDept}(x, y)\}$, e.g., the pair $\langle \mathsf{Mary}, \mathsf{John} \rangle$, which is not witnessed by any source data. Nonetheless, there are facts at the global level, as for example $\mathit{EmployeeDept}(\mathsf{Mary}, \mathsf{D1})$, that would be entailed by the system even in the absence of the inconsistency described above. Therefore, it seems reasonable to assume that the set of "significant" certain answers to our query is the set $\{\langle \mathsf{Mary}, \mathsf{D1} \rangle\}$, rather than the set of all pairs of elements of Γ.

To the aim of overcoming the problems illustrated above, we characterize the semantics of a data integration system $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ with respect to a source instance $\mathcal{D}$ in terms of those interpretations over Δ of the symbols in $\mathcal{A}_\mathcal{G}$ that:

1. satisfy the global schema $\mathcal{G}$;
2. satisfy *as much as possible* the mapping assertions in $\mathcal{M}$ with respect to the source instance $\mathcal{D}$.

In other words, under this assumption, the knowledge expressed by $\mathcal{G}$ is considered more reliable than the knowledge represented by the information retrieved at the data sources through the mapping assertions.

In order to determine the precise meaning of "satisfying as much as possible" the mapping with respect to a source instance $\mathcal{D}$, we define preference orders over the models of $\mathcal{G}$.

Let $\mathcal{U}_\Delta$ be the set of interpretations of $\mathcal{A}_\mathcal{G}$ over Δ, and let $\succeq$ be a reflexive and transitive binary relation defined over $\mathcal{U}_\Delta \times \mathcal{U}_\Delta$ that depends on the mapping $\mathcal{M}$ and the source database $\mathcal{D}$. The relation $\succeq$ induces a preference order over the global interpretations of the system. More precisely, given two interpretations $\mathcal{W}, \mathcal{W}'$ of $\mathcal{A}_\mathcal{G}$, we say that $\mathcal{W}'$ is $\succeq$*-preferred to* $\mathcal{W}$ if $\mathcal{W}' \succeq \mathcal{W}$ and $\mathcal{W} \not\succeq \mathcal{W}'$.

Then, we are ready to generalize Definition 1 and give a new notion of model for an integration system $\mathcal{I}$ with respect to a source model $\mathcal{D}$, which corresponds to the notion of maximal element in the preference order defined above.

Definition 3. Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a data integration system, let $\mathcal{D}$ be a source model for $\mathcal{I}$, let $\mathcal{W}$ be an interpretation over $\mathcal{A}_\mathcal{G}$, and let $\succeq$ be a reflexive and transitive binary relation defined over $\mathcal{U}_\Delta \times \mathcal{U}_\Delta$ that depends on $\mathcal{M}$ and $\mathcal{D}$. We say that $\mathcal{W}$ is an $\succeq$*-model for* $(\mathcal{I}, \mathcal{D})$ if $\mathcal{W}$ is a model for $\mathcal{G}$, and for each model $\mathcal{W}'$ for $\mathcal{G}$, $\mathcal{W}'$ is not $\succeq$-preferred to $\mathcal{W}$.

The previous definition allows for defining a new semantics for a data integration system $\mathcal{I}$ with respect to a source database $\mathcal{D}$. In particular, we define

$$\mathit{consSem}(\succeq, \mathcal{I}, \mathcal{D}) = \{\mathcal{W} \mid \mathcal{W} \text{ is a } \succeq\text{-model for } (\mathcal{I}, \mathcal{D})\}.$$

Now, we instantiate the above general semantics by defining three distinct preference relations over the interpretations of $\mathcal{A}_\mathcal{G}$. Informally, we consider as intended models of the integration system those interpretations that satisfy $\mathcal{G}$ and satisfy as much as possible a set of first-order sentences that constitutes the "image of the mapping assertions" with respect to $\mathcal{D}$. More precisely, we define three different criteria for comparing two interpretations, based on the different relevance we attribute to sound and complete mapping assertions,

i.e., assertions of the form $q_S \sqsubseteq q_G$ and $q_G \sqsubseteq q_S$, respectively. This approach gives rise to three different semantics:

1. *consSem*$_S$, where sound mapping assertions are more relevant than complete mapping assertions;
2. *consSem*$_C$, where complete mapping assertions are more relevant than sound mapping assertions;
3. *consSem*, where all mapping assertions have the same relevance.

To formalize the above ideas, we first define the notions of "image" of the mapping $\mathcal{M}$ with respect to a model $\mathcal{D}$ of the sources as a set of first-order sentences. In the following definition, $q(t)$ indicates the FOL sentence obtained from the open formula q by replacing its free variables with the constants in t, i.e., if $t = \langle t_1, \ldots, t_n \rangle$ and $\{x_1, \ldots, x_n\}$ are the free variables of q, $x_i = t_i$ for each $1 \leqslant i \leqslant n$.

Definition 4. Given a data integration system $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ and a source model $\mathcal{D}$ for $\mathcal{I}$, we define *S-Image*$(\mathcal{M}, \mathcal{D})$, *C-Image*$(\mathcal{M}, \mathcal{D})$, and *Image*$(\mathcal{M}, \mathcal{D})$ as follows:

$$\begin{aligned}
&\textit{S-Image}(\mathcal{M}, \mathcal{D}) = \big\{ q_G(t) \mid q_S \sqsubseteq q_G \in \mathcal{M} \text{ and } t \in q_S^{\mathcal{D}} \big\},\\
&\textit{C-Image}(\mathcal{M}, \mathcal{D}) = \big\{ \neg q_G(t) \mid q_G \sqsubseteq q_S \in \mathcal{M} \text{ and } t \text{ is a tuple of } \Gamma \text{ and } t \notin q_S^{\mathcal{D}} \big\},\\
&\textit{Image}(\mathcal{M}, \mathcal{D}) = \textit{S-Image}(\mathcal{M}, \mathcal{D}) \cup \textit{C-Image}(\mathcal{M}, \mathcal{D}).
\end{aligned}$$

Intuitively, *S-Image*$(\mathcal{M}, \mathcal{D})$ represents the "image" of the sound mapping assertions with respect to $\mathcal{D}$, while *C-Image*$(\mathcal{M}, \mathcal{D})$ represents the image of the complete mapping assertions with respect to $\mathcal{D}$, and *Image*$(\mathcal{M}, \mathcal{D})$ is the image of all mapping assertions with respect to $\mathcal{D}$.

Example 1 (*contd.*). In our ongoing example, we have that

$$\begin{aligned}
\textit{S-Image}(\mathcal{M}_0, \mathcal{D}_0') = \big\{ &\textit{EmployeeDept}(\mathsf{Mary}, \mathsf{D1}) \wedge \textit{DeptLocation}(\mathsf{D1}, \mathsf{New\ York}),\\
&\textit{EmployeeDept}(\mathsf{John}, \mathsf{D2}) \wedge \textit{DeptLocation}(\mathsf{D2}, \mathsf{New\ York}),\\
&\textit{DeptDirector}(\mathsf{D1}, \mathsf{John}) \big\}, \text{ and}\\
\textit{C-Image}(\mathcal{M}_0, \mathcal{D}_0') = \big\{ &\neg \textit{DeptDirector}(\alpha, \beta) \mid \alpha, \beta \in \Gamma \text{ and } \alpha \neq \mathsf{D1} \text{ or } \beta \neq \mathsf{John} \big\}.
\end{aligned}$$

Then, given an interpretation $\mathcal{W}$ of the elements in $\mathcal{A}_G$, we define *SatIm*$(\mathcal{W}, \mathcal{M}, \mathcal{D})$ as the portion of the image of $\mathcal{M}$ with respect to $\mathcal{D}$ satisfied by $\mathcal{W}$. More precisely:

Definition 5. Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a data integration system, let $\mathcal{D}$ be a source model for $\mathcal{I}$, and let $\mathcal{W}$ be a global interpretation of $\mathcal{I}$. We define:

$$\begin{aligned}
&\textit{S-SatIm}(\mathcal{W}, \mathcal{M}, \mathcal{D}) = \big\{ \varphi \mid \varphi \in \textit{S-Image}(\mathcal{M}, \mathcal{D}) \text{ and } \mathcal{W} \models \varphi \big\},\\
&\textit{C-SatIm}(\mathcal{W}, \mathcal{M}, \mathcal{D}) = \big\{ \varphi \mid \varphi \in \textit{C-Image}(\mathcal{M}, \mathcal{D}) \text{ and } \mathcal{W} \models \varphi \big\},\\
&\textit{SatIm}(\mathcal{W}, \mathcal{M}, \mathcal{D}) = \textit{S-SatIm}(\mathcal{W}, \mathcal{M}, \mathcal{D}) \cup \textit{C-SatIm}(\mathcal{W}, \mathcal{M}, \mathcal{D}).
\end{aligned}$$

Based on the above notions of image of the mapping with respect to a source instance, we are now ready to define three partial orders, relying on set containment, over the global interpretations of a data integration system.

Definition 6. Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a data integration system, let $\mathcal{D}$ be a source model for $\mathcal{I}$, and let $\mathcal{W}, \mathcal{W}'$ be two global interpretations for $\mathcal{I}$. We define the relations $\succeq^S_{(\mathcal{M},\mathcal{D})}$, $\succeq^C_{(\mathcal{M},\mathcal{D})}$, $\succeq_{(\mathcal{M},\mathcal{D})}$ as follows:

1. $\mathcal{W}' \succeq^S_{(\mathcal{M},\mathcal{D})} \mathcal{W}$ if one of the following conditions holds:
   (a) $S\text{-}SatIm(\mathcal{W}', \mathcal{M}, \mathcal{D}) \supset S\text{-}SatIm(\mathcal{W}, \mathcal{M}, \mathcal{D})$;
   (b) $S\text{-}SatIm(\mathcal{W}', \mathcal{M}, \mathcal{D}) = S\text{-}SatIm(\mathcal{W}, \mathcal{M}, \mathcal{D})$ and $C\text{-}SatIm(\mathcal{W}', \mathcal{M}, \mathcal{D}) \supset C\text{-}SatIm(\mathcal{W}, \mathcal{M}, \mathcal{D})$.
2. $\mathcal{W}' \succeq^C_{(\mathcal{M},\mathcal{D})} \mathcal{W}$ if one of the following conditions holds:
   (a) $C\text{-}SatIm(\mathcal{W}', \mathcal{M}, \mathcal{D}) \supset C\text{-}SatIm(\mathcal{W}, \mathcal{M}, \mathcal{D})$;
   (b) $C\text{-}SatIm(\mathcal{W}', \mathcal{M}, \mathcal{D}) = C\text{-}SatIm(\mathcal{W}, \mathcal{M}, \mathcal{D})$ and $S\text{-}SatIm(\mathcal{W}', \mathcal{M}, \mathcal{D}) \supset S\text{-}SatIm(\mathcal{W}, \mathcal{M}, \mathcal{D})$.
3. $\mathcal{W}' \succeq_{(\mathcal{M},\mathcal{D})} \mathcal{W}$ if $SatIm(\mathcal{W}', \mathcal{M}, \mathcal{D}) \supset SatIm(\mathcal{W}, \mathcal{M}, \mathcal{D})$.

The previous definition allows for specializing the $consSem(\succeq, \mathcal{I}, \mathcal{D})$, and defining the semantics for each of the above partial orders. In particular:

$$consSem_S(\mathcal{I}, \mathcal{D}) = \{\mathcal{W} \mid \mathcal{W} \text{ is a } \succeq^S_{(\mathcal{M},\mathcal{D})}\text{-model for } (\mathcal{I}, \mathcal{D})\},$$
$$consSem_C(\mathcal{I}, \mathcal{D}) = \{\mathcal{W} \mid \mathcal{W} \text{ is a } \succeq^C_{(\mathcal{M},\mathcal{D})}\text{-model for } (\mathcal{I}, \mathcal{D})\},$$
$$consSem(\mathcal{I}, \mathcal{D}) = \{\mathcal{W} \mid \mathcal{W} \text{ is a } \succeq_{(\mathcal{M},\mathcal{D})}\text{-model for } (\mathcal{I}, \mathcal{D})\}.$$

Example 2. Consider the data integration system $\mathcal{I}_1 = \langle \mathcal{G}_1, \mathcal{S}_1, \mathcal{M}_1 \rangle$, such that the global alphabet $\mathcal{A}_{\mathcal{G}_1}$ contains the binary relation symbol *relative*, which indicates pairs of relatives, and that the following FOL sentence is specified over $\mathcal{A}_{\mathcal{G}_1}$,

$$\forall x, y.relative(x, y) \supset relative(y, x),$$

stating that if x is a relative of y also the converse holds. Assume now that $\mathcal{S}_1$ contains the binary relation symbol s and that the mapping $\mathcal{M}_1$ is as follows,

$$relative(x, y) \sqsubseteq s(x, y),$$
$$s(x, y) \sqsubseteq relative(x, y).$$

Then, let $\mathcal{D}_1 = \{s(\mathsf{Albert}, \mathsf{Ann})\}$ be a source model for $\mathcal{I}_1$. It is easy to see that

$$S\text{-}Image(\mathcal{M}_1, \mathcal{D}_1) = \{relative(\mathsf{Albert}, \mathsf{Ann})\},$$
$$C\text{-}Image(\mathcal{M}_1, \mathcal{D}_1) = \{\neg relative(\alpha, \beta) \mid \alpha, \beta \in \Gamma \text{ and } \alpha \neq \mathsf{Albert} \text{ or } \beta \neq \mathsf{Ann}\}.$$

Therefore, we have that

$$consSem_S(\mathcal{I}_1, \mathcal{D}_1) = \{\mathcal{W} \mid \mathcal{W} \models \mathcal{G}_1 \text{ and } \mathcal{W} \supseteq \{relative(\mathsf{Albert}, \mathsf{Ann}), relative(\mathsf{Ann}, \mathsf{Albert})\}\},$$

$$consSem_C(\mathcal{I}_1, \mathcal{D}_1) = \{\emptyset\}, \quad \text{and}$$
$$consSem(\mathcal{I}_1, \mathcal{D}_1) = consSem_S(\mathcal{I}_1, \mathcal{D}_1) \cup consSem_C(\mathcal{I}_1, \mathcal{D}_1).$$

Finally, we are able to define the notion of certain answers in the new semantics.

Definition 7. Let $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ be a data integration system, let $\mathcal{D}$ be a source model for $\mathcal{I}$, and let q be a query over $\mathcal{G}$. Then:

$$consAns_S(q, \mathcal{I}, \mathcal{D}) = \{t \mid t \in q^{\mathcal{W}} \text{ for each } \mathcal{W} \in consSem_S\},$$
$$consAns_C(q, \mathcal{I}, \mathcal{D}) = \{t \mid t \in q^{\mathcal{W}} \text{ for each } \mathcal{W} \in consSem_C\},$$
$$consAns(q, \mathcal{I}, \mathcal{D}) = \{t \mid t \in q^{\mathcal{W}} \text{ for each } \mathcal{W} \in consSem\}.$$

Example 1 (*contd.*). Let us first enumerate the sentences of $S\text{-}Image(\mathcal{M}_0, \mathcal{D}'_0)$ as follows:

1. $EmployeeDept(\mathsf{Mary}, \mathsf{D1}) \wedge DeptLocation(\mathsf{D1}, \mathsf{New\ York})$,
2. $EmployeeDept(\mathsf{John}, \mathsf{D2}) \wedge DeptLocation(\mathsf{D2}, \mathsf{New\ York})$,
3. $DeptDirector(\mathsf{D1}, \mathsf{John})$.

Then, we have that $consSem(\mathcal{I}_0, \mathcal{D}'_0)$ contains all global interpretations $\mathcal{W}_0$ of $\mathcal{I}_0$ that satisfy either sentences 1 and 2 or sentences 1 and 3. Indeed, if $\mathcal{W}_0$ satisfied both sentences 2 and 3, the facts $EmployeeDept(\mathsf{John}, \mathsf{D2})$ and $DeptDirector(\mathsf{D1}, \mathsf{John})$ would hold in $\mathcal{W}$, and hence also the fact $EmployeeDept(\mathsf{John}, \mathsf{D1})$ would hold, since in $\mathcal{G}_0$ a director of a department is also an employee of the same department. Thus, $\mathcal{W}_0$ would violate the sentence in $\mathcal{G}_0$ stating that each employee works in only one department. On the other hand, $\mathcal{W}_0$ cannot satisfy only one sentence in $S\text{-}Image(\mathcal{M}_0, \mathcal{D}'_0)$ or any sentence in $S\text{-}Image(\mathcal{M}_0, \mathcal{D}'_0)$, since in such a way it would not be maximal with respect to the $\succeq_{(\mathcal{M},\mathcal{D})}$-preference ordering.

Notice that, for the query $q = \{x, y \mid EmployeeDept(x, y)\}$ we have that $consAns(q, \mathcal{I}_0, \mathcal{D}'_0) = \{\langle \mathsf{Mary}, \mathsf{D1} \rangle\}$.

We point out that the semantics *consSem* (and also $consSem_S$ and $consSem_C$) defined above has an important property: for each integration system $\mathcal{I}$ and source instance $\mathcal{D}$, if $sem(\mathcal{I}, \mathcal{D}) \neq \emptyset$ then $consSem(\mathcal{I}, \mathcal{D}) = sem(\mathcal{I}, \mathcal{D})$ (the same equality holds both for $consSem_S$ and $consSem_C$). In this sense, such semantics can be considered as "conservative extensions" of the classical semantics *sem*, since they provide a different meaning to a data integration system only in the presence of inconsistency (i.e., only when $sem(\mathcal{I}, \mathcal{D}) = \emptyset$).

As a concluding remark, observe that to specialize the above semantics in order to adopt a cardinality-based preference criterion for models, rather than a set-containment-based one, it suffices to suitably modify Definition 6, comparing the cardinality of the sets $SatIm(\mathcal{W}, \mathcal{M}, \mathcal{D})$ and $SatIm(\mathcal{W}', \mathcal{M}, \mathcal{D})$ instead of their extension. As we shall see in the next section, such a quantitative approach has been proposed in the literature (e.g., [41]).

4. Comparison with current proposals

The framework for data integration presented in the previous sections is very general, in terms of (i) global schema (first-order theories), (ii) mapping assertions (generalization of GAV and LAV, first-order queries), (iii) semantics. In this section, we briefly survey the main studies both in the area of data integration and in the field of *inconsistent databases*, which studies the problem of computing answers to databases in which data violate integrity constraints, and we show that our framework is able to capture all the logic-based approaches to data integration and to inconsistent databases proposed in the literature. Such an analysis allows for a better understanding of the different semantic nature of the existing proposals.

4.1. Relationship with belief revision and update

First of all, we point out that the problem of reasoning with inconsistent databases is closely related to the studies in *belief revision and update* [3,35,45]. This area of Artificial Intelligence studies the problem of integrating new information with previous knowledge. In general, the problem is studied in a logical framework, in which the new information is a logical formula f and the previous knowledge is a logical theory (also called knowledge base) T. Of course, f may in general be inconsistent with T. The *revised* (or *updated*) knowledge base is denoted as $T \circ f$, and several semantics have been proposed for the operator $\circ$. The semantics for belief revision can be divided into *revision* semantics, when the new information f is interpreted as a modification of the knowledge about the world, and *update* semantics, when f reflects a change in the world.

The problem of reasoning about a data integration system $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$, whose data at the sources $\mathcal{D}$ may be inconsistent with respect to the global schema and the mapping, can be actually seen as a problem of belief revision. In fact, with respect to the above illustrated knowledge base revision framework, if we consider the source instance $\mathcal{D}$ and the mapping specification $\mathcal{M}$ as the initial knowledge base T, and the global schema $\mathcal{G}$ as the new information f, then the problem of deciding whether a tuple t is in the answer set of a query q with respect to the system $\mathcal{I}$ and the source instance $\mathcal{D}$ corresponds to the belief revision problem $(\mathcal{D} \cup \mathcal{M}) \circ \mathcal{G} \models q(t)$.

Based on such a correspondence, the studies in belief revision appear very relevant for the field of data integration: indeed, our framework can be seen in principle as the application of a semantics for belief revision/update in a particular class of logical theories (for a detailed definition of some of the most important belief revision/update semantics see e.g. [25]).

However, due to the structure of a data integration system, the kind of theories that must be revised/updated have a very special form. Specifically, in a data integration architecture, the mapping assertions, which are sentences of a very particular form (implication of first-order queries), provide the only connection that exists between data at the sources, which are part of the initial knowledge, and the global schema, which represents the revised knowledge. Hence, mapping assertions constitute the crucial part of the theory in the revision/update process in data integration. Due to the form of such assertions, it is possible to define a semantic treatment of revision/update which is specialized for this particular kind

of sentences. This is precisely what is generally done in the data integration literature, and what we have proposed in our framework: preferred model of the revised/updated theory must maximize satisfaction of the mapping assertions.

On the other hand, even in the context of database update/revision, which is the closest to the data integration setting, the concept of mapping is missing, which in general makes it hard to provide a detailed comparison of the semantic approaches presented in this paper with the literature on database update [35,45]. However, belief revision/update in a typical database setting is considered by the literature on *inconsistent databases*, which we briefly survey in the following.

4.2. Consistent query answering in inconsistent databases

We now briefly survey the main existing approaches to inconsistent databases. We start by pointing out that the single database setting, that is the one that is studied in the field of inconsistent databases, can be seen as a very special case of a data integration scenario. Indeed, a relational schema $\mathcal{RS}$ corresponds to the global schema $\mathcal{G}$ of a data integration system $\mathcal{I} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M} \rangle$ in which relation predicates in $\mathcal{G}$ are in a one-to-one correspondence with relation predicates in $\mathcal{S}$. More precisely, if $g_1/h_1, \ldots, g_n/h_n$ are the global relations, where with g_i/h_i we indicate that h_i is the arity of g_i, then the source relations are $s_1/h_1, \ldots, s_n/h_n$, and the mapping is given by the n one-to-one assertions

$$\{X_1, \ldots, X_{h_i} \mid g_i(X_1, \ldots, X_{h_i})\} \sqsubseteq \{X_1, \ldots, X_{h_i} \mid s_i(X_1, \ldots, X_{h_i})\}$$

for each i, $1 \leqslant i \leqslant n$ in the case of sound mapping, while the assertions have the form

$$\{X_1, \ldots, X_{h_i} \mid s_i(X_1, \ldots, X_{h_i})\} \sqsubseteq \{X_1, \ldots, X_{h_i} \mid g_i(X_1, \ldots, X_{h_i})\}$$

in the case of complete mapping (both kinds of assertions are expressed in the case of an exact mapping). With this notion in place, we can review the works in inconsistent databases by comparing them with our data integration framework.

Arenas et al. define in [4] a semantics for handling databases in which data are inconsistent with respect to a set of integrity constraints, and an algorithm for computing certain answers (called *consistent answers*) to user queries under such a semantics. The query answering method is proved to be sound and complete only for the class of universally quantified binary constraints, i.e., non-existential FOL sentences of a particular form that involve two database relations. In [5], the same authors propose a new method based on the use of *logic rules with exceptions* that can handle arbitrary universally quantified constraints. The semantics underlying the notion of consistent query answers both in [4] and in [5] is defined on a set-containment ordering between databases. It turns out that this approach corresponds in our framework to the case of an exact, one-to-one mapping and to the *consSem* semantics.

Greco et al. propose in [29] a technique to deal with inconsistencies that is based on the reformulation of integrity constraints into a disjunctive Datalog program with two different forms of negation: negation as failure and classical negation. Such a program can be used both to repair databases, i.e., modify the data in the databases in order to satisfy integrity constraints, and to compute certain answers to queries. The technique is proved to be sound and complete for universally quantified constraints. Also in this case, such an approach is

captured in our framework by adopting an exact, one-to-one mapping and by the notion of *consSem*.

In [27], Fagin et al. propose a framework for updating theories and logical databases (i.e., databases obtained by giving priorities to sentences in the databases) that can be extended also to the case of updating views. The semantics proposed in that paper is based on a particular set-containment based ordering between theories that "accomplish" an update to an original theory. More precisely, a theory T_1 accomplishes an insertion of a fact σ into T if $\sigma \in T_1$, and accomplishes a deletion of σ if σ is not a logical consequence of T_1. Then, a theory T_1 accomplishes an update u to T with a *smaller change than* T_2, and thus is preferred to T_2, if both T_1 and T_2 accomplish u, and either: (1) the set of facts deleted from T to obtain T_1 is contained in the set of facts deleted from T to obtain T_2 (notice that no condition on the added facts is imposed); or (2) the two sets of deleted facts described above coincide, but the set of facts added to T to obtain T_1 is contained in the analogous set needed to obtain T_2 from T. It is easy to verify that this approach corresponds in our framework to an exact one-to-one mapping and to the notion of $\mathit{consSem}_S$.

Moreover, a different semantics for database repairing has been considered by Chomicki et al. in [21,22]. Specifically, in such works a semantics is defined in which only elimination of tuples is allowed; therefore, the problem of dealing with infinite models is not addressed. Then, a preference order over the database repairs is defined, in such a way that only minimal repairs (in terms of set containment) are considered. Hence, the semantics is a "maximal complete" one, in the sense that only maximal consistent subsets of the database instance are considered as repairs of such an instance. In [22] the authors establish complexity results for query answering under such a semantics in the presence of denial constraints [2], while in [21] also inclusion dependencies [2] are considered. This approach corresponds in our framework to an exact one-to-one mapping and to the notion of $\mathit{consSem}_C$. Although in a different formal framework, the same semantic approach is also considered by Baral et al. in [6].

A cardinality-based approach is pursued by Lin et al. in [41], where the authors describe an operator for *merging databases* under constraints which allows for obtaining a maximal amount of information from each database by means of a majority criterion used in case of conflict. Notice that, differently from all the other studies mentioned above, this approach relies on a cardinality-based ordering between databases (rather than a set-containment-based ordering). However, our general framework is able to capture this approach: specifically, the semantic principle adopted in [41] is exactly captured by Definition 3 under the following relation $\succeq$: given the mapping $\mathcal{M}$ and a source model $\mathcal{D}$, $\mathcal{W}' \succeq \mathcal{W}$ iff $\mathit{dist}(\mathcal{W}', \mathit{Image}(\mathcal{M}, \mathcal{D})) \leqslant \mathit{dist}(\mathcal{W}, \mathit{Image}(\mathcal{M}, \mathcal{D}))$, i.e., the interpretations are ordered according to their "distance" from the theory $\mathit{Image}(\mathcal{M}, \mathcal{D})$, where

$$\mathit{dist}\big(\mathcal{W}, \mathit{Image}(\mathcal{M}, \mathcal{D})\big) = \min_{\mathcal{W}_i \models \mathit{Image}(\mathcal{M}, \mathcal{D})} \big(|\mathcal{W} - \mathcal{W}_i| + |\mathcal{W}_i - \mathcal{W}|\big),$$

i.e., the distance between an interpretation $\mathcal{W}$ and $\mathit{Image}(\mathcal{M}, \mathcal{D})$ is the minimum distance between $\mathcal{W}$ and any model of $\mathit{Image}(\mathcal{M}, \mathcal{D})$, where the distance between two interpretations is measured in terms of the cardinality of the symmetric difference of the interpretations.

Finally, Calì et al. [14] present three different semantics for inconsistent databases, called respectively *loosely-sound, loosely-exact, loosely-complete* semantics. They all cor-

respond to instances of the semantics *consSem* of our framework, where the one-to-one mapping is defined respectively through sound, exact, and complete mapping assertions.

4.3. Data integration

In the field of data integration, most of the logic-based approaches have adopted a classical first-order semantics ([31,38,44] provide a complete picture of the main works in this area). In particular, all the approaches that use LAV mapping assertions adopt a sound semantics for the mapping (see e.g. [1,7,17,24,30,43]), while the studies concerning GAV mapping assertions have in general interpreted the mapping as exact (e.g., [9,20]). A notable exception for GAV is [13], where a sound assumption on the mapping assertions is adopted.

Only recently the problem of dealing with inconsistent data has been taken into account in logic-based data integration settings. In particular, data inconsistency in a LAV scenario has been studied in [10] and [11]. The semantics proposed in [10] and [11] turns out to be different from each of the semantics proposed in our framework. Indeed, while our proposal focuses on the mapping and define a suitable relaxation of it in the presence of inconsistency, [10,11] characterize the semantics in terms of the repairs of the different global databases that can be obtained by populating the global schema according to the LAV mapping. More specifically, [10,11] assume that the mapping is sound, and consider the set $\min(\mathcal{G})$ of the minimal (with respect to set inclusion) global databases that satisfy the mapping with respect to the source instance. Then, the models of the system, called *repairs*, are the global databases consistent with the constraints on the global schema that are minimal with respect to $\leqslant_{\mathcal{DB}}$ for some $\mathcal{DB} \in \min(\mathcal{G})$, where $\mathcal{B} \leqslant_{\mathcal{DB}} \mathcal{B}'$ if $\triangle(\mathcal{B}, \mathcal{DB}) \subseteq \triangle(\mathcal{B}', \mathcal{DB})$, where in turn $\triangle(X, Y)$ indicates the symmetric difference between X and Y. In this semantics, even if the mapping is assumed to be sound, the repairs are computed on each database in $\min(\mathcal{G})$, as if the retrieved data were exact. Therefore, the semantics is not "mapping-centered" as in our framework. Moreover, the repair semantics can be different from the first-order semantics even when the latter is not empty.

Finally, in [15] the framework based on the *loosely-sound* semantics, introduced for inconsistent databases in [14], is extended to the data integration setting. More precisely, relational global schemas and GAV mapping assertions are considered. This corresponds in our framework to the *consSem* semantics under sound mapping assertions.

We summarize the analysis described above in the table reported in Table 1, which presents a classification of the literature considered in this section. The table has four main rows, which represent the four semantics we have defined in our framework, and three columns, one for each possible kind of mapping. In each cell of the table we have reported the approaches that adopt the corresponding combination of semantics and mapping. As it is immediate to see in the table, almost all the mentioned studies in inconsistent databases can be considered as data integration approaches adopting an exact mapping and a "symmetric" semantics *consSem*, while the main approaches to data integration adopt a sound mapping and the classical first-order semantics *sem*.

Table 1
Classification of the approaches considered in the paper

	Sound mapping ($q_S \sqsubseteq q_G$)	Exact mapping ($q_S \sqsubseteq q_G$ and $q_G \sqsubseteq q_S$)	Complete mapping ($q_G \sqsubseteq q_S$)
sem	Abiteboul et al. [1] Duschka et al. [24] Calvanese et al. [17]	Chawathe et al. [20] Bergamaschi et al. [9]	
consSem	Calì et al. [14] (loosely-sound)	Bry [12] Arenas et al. [4,5] Greco et al. [29] Calì et al. [14] (loosely-exact)	Calì et al. [14] (loosely-complete)
	Calì et al. [15]	Lin et al. [41] (card.)	
$consSem_S$		Fagin et al. [27]	
$consSem_C$		Chomicki et al. [21] Baral et al. [6]	

5. Conclusions

In this paper, we have studied the problem of data integration in the general setting in which data at the sources may result inconsistent with respect to the knowledge modeled by the integration system. In particular, we have defined a comprehensive formal framework which is able to capture the main logical approaches to data integration proposed so far in the literature, and we have compared different proposals on the basis of such a framework. Moreover, our "mapping-centered" semantics has allowed us to highlight the crucial role played by the mapping in data integration systems, and to amplify the structural differences of the data integration scenario with respect to the setting of a single database.

In the present work, whose focus was on the semantic aspects related to data integration, we have not considered the crucial problem of *query processing* in data integration systems under the different semantics proposed in our framework. In this respect, we remark that the complexity of the task of computing the answers to queries is not only influenced by the criterion chosen for dealing with inconsistency, but also heavily depends on the expressiveness of the formalism used for modeling the system. More precisely, the complexity of query processing depends on all the aspects listed at the beginning of Section 2, and in particular: (i) the user query language; (ii) the language for expressing the mapping; (iii) the formalism for expressing the global schema (e.g., the form of the integrity constraints that can be expressed over the global schema). The first studies concerning the decidability and complexity of query processing in such a rich and complex setting have appeared only recently (e.g., [4,14,15,18,22,29]).

The formal framework for data integration presented in this paper may be extended in several directions. For instance, it should be worth addressing the presence of more complex forms of data sources in the integration system. Moreover, it would be very interesting to generalize our approach to more involved information integration scenarios, e.g., data-exchange [26] and peer-to-peer systems [33], in which the assumption of a global information schema is unrealistic.

Acknowledgements

This research has been partially supported by the projects INFOMIX (IST-2001-33570), SEWASIE (IST-2001-34825) and INTEROP Network of Excellence (IST-508011) funded by the EU, by the project "Società dell'Informazione" subproject SP1 "Reti Internet: Efficienza, Integrazione e Sicurezza" funded by MIUR – Fondo Speciale per lo Sviluppo della Ricerca di Interesse Strategico, by project HYPER, funded by IBM through a Shared University Research (SUR) Award grant, and by project MAIS (Multichannel Adaptive information Systems), supported by MIUR under FIRB (Fondo Italiano per la Ricerca di Base).

References

[1] S. Abiteboul, O. Duschka, Complexity of answering queries using materialized views, in: Proceedings of the 17th ACM SIGACT SIGMOD SIGART Symposium on Principles of Database Systems (PODS'98), 1998, pp. 254–265.
[2] S. Abiteboul, R. Hull, V. Vianu, Foundations of Databases, Addison-Wesley Publ. Co., Reading, MA, 1995.
[3] C.E. Alchourrón, P. Gärdenfors, D. Makinson, On the logic of theory change: Partial meet contraction and revision functions, J. Symbolic Logic 50 (1985) 510–530.
[4] M. Arenas, L.E. Bertossi, J. Chomicki, Consistent query answers in inconsistent databases, in: Proceedings of the 11th ACM SIGACT SIGMOD SIGART Symposium on Principles of Database Systems (PODS'99), 1999, pp. 68–79.
[5] M. Arenas, L.E. Bertossi, J. Chomicki, Specifying and querying database repairs using logic programs with exceptions, in: Proceedings of the 4th International Conference on Flexible Query Answering Systems (FQAS 2000), Springer, 2000, pp. 27–41.
[6] C. Baral, S. Kraus, J. Minker, V.S. Subrahamanian, Combining knowledge bases consisting of first-order analysis, Comput. Intelligence 8 (1992) 45–71.
[7] C. Beeri, A.Y. Levy, M.-C. Rousset, Rewriting queries using views in description logics, in: Proceedings of the 16th ACM SIGACT SIGMOD SIGART Symposium on Principles of Database Systems (PODS'97), 1997, pp. 99–108.
[8] D. Beneventano, S. Bergamaschi, S. Castano, A. Corni, R. Guidetti, G. Malvezzi, M. Melchiori, M. Vincini, Information integration: the MOMIS project demonstration, in: Proceedings of the 26th International Conference on Very Large Data Bases (VLDB 2000), 2000.
[9] S. Bergamaschi, S. Castano, M. Vincini, D. Beneventano, Semantic integration of heterogeneous information sources, Data Knowledge Engrg. 36 (3) (2001) 215–249.
[10] L. Bertossi, J. Chomicki, A. Cortes, C. Gutierrez, Consistent answers from integrated data sources, in: Proceedings of the 6th International Conference on Flexible Query Answering Systems (FQAS 2002), 2002, pp. 71–85.
[11] L. Bravo, L. Bertossi, Logic programming for consistently querying data integration systems, in: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), 2003, pp. 10–15.
[12] F. Bry, Query answering in information systems with integrity constraints, in: IFIP WG 11.5 Working Conference on Integrity and Control in Information System, Chapman & Hall, 1997.
[13] A. Calì, D. Calvanese, G. De Giacomo, M. Lenzerini, Data integration under integrity constraints, Inform. Syst. 29 (2004) 147–163.
[14] A. Calì, D. Lembo, R. Rosati, On the decidability and complexity of query answering over inconsistent and incomplete databases, in: Proceedings of the 22nd ACM SIGACT SIGMOD SIGART Symposium on Principles of Database Systems (PODS 2003), 2003, pp. 260–271.
[15] A. Calì, D. Lembo, R. Rosati, Query rewriting and answering under constraints in data integration systems, in: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), 2003, pp. 16–21.

[16] D. Calvanese, G. De Giacomo, M. Lenzerini, Answering queries using views over description logics knowledge bases, in: Proceedings of the 17th National Conference on Artificial Intelligence (AAAI 2000), 2000, pp. 386–391.
[17] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, R. Rosati, Data integration in data warehousing, Internat. J. Cooperat. Inform. Syst. 10 (3) (2001) 237–271.
[18] D. Calvanese, R. Rosati, Answering recursive queries under keys and foreign keys is undecidable, in: Proceedings of the 10th International Workshop on Knowledge Representation meets Databases (KRDB 2003), 2003;
CEUR Electronic Workshop Proceedings http://ceur-ws.org/Vol-79/.
[19] W.A. Carnielli, J. Marcos, A taxonomy of C-systems, in: Paraconsistency—the Logical Way to the Inconsistent, in: Lecture Notes in Pure and Applied Mathematics, 2001, pp. 1–94.
[20] S.S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J.D. Ullman, J. Widom, The TSIMMIS project: Integration of heterogeneous information sources, in: Proc. of the 10th Meeting of the Information Processing Society of Japan (IPSJ'94), 1994, pp. 7–18.
[21] J. Chomicki, J. Marcinkowski, Minimal-change integrity maintenance using tuple deletions, Technical Report cs.DB/0212004 v1, arXiv.org e-Print archive, December 2002. Available at http://arxiv.org/abs/cs/0212004.
[22] J. Chomicki, J. Marcinkowski, On the computational complexity of consistent query answers. Technical Report cs.DB/0204010 v1, arXiv.org e-Print archive, April 2002, Available at http://arxiv.org/abs/cs/0204010.
[23] N.C.A. da Costa, On the theory of inconsistent formal systems, Notre Dame J. Formal Logic 15 (1974) 497–510.
[24] O.M. Duschka, M.R. Genesereth, A.Y. Levy, Recursive query plans for data integration, J. Logic Programm. 43 (1) (2000) 49–73.
[25] T. Eiter, G. Gottlob, On the complexity of propositional knowledge base revision, updates and counterfactuals, Artificial Intelligence 57 (1992) 227–270.
[26] R. Fagin, P.G. Kolaitis, R.J. Miller, L. Popa, Data exchange: Semantics and query answering, in: Proceedings of the 9th International Conference on Database Theory (ICDT 2003), 2003, pp. 207–224.
[27] R. Fagin, J.D. Ullman, M.Y. Vardi, On the semantics of updates in databases, in: Proceedings of the 2nd ACM SIGACT SIGMOD Symposium on Principles of Database Systems (PODS'83), 1983, pp. 352–365.
[28] M.R. Genereseth, A.M. Keller, O.M. Duschka, Infomaster: An information integration system, in: ACM SIGMOD International Conference on Management of Data, 1997.
[29] G. Greco, S. Greco, E. Zumpano, A logical framework for querying and repairing inconsistent databases, IEEE Trans. Knowledge Data Engrg. 15 (6) (2003) 1389–1408.
[30] J. Gryz, Query rewriting using views in the presence of functional and inclusion dependencies, Inform. Syst. 24 (7) (1999) 597–612.
[31] A.Y. Halevy, Theory of answering queries using views, SIGMOD Record 29 (4) (2000) 40–47.
[32] A.Y. Halevy, Answering queries using views: A survey, Very Large Database J. 10 (4) (2001) 270–294.
[33] A.Y. Halevy, G.I. Zachary, D. Suciu, I. Tatarinov, Schema mediation in peer data management systems, in: Proceedings of the 19th IEEE International Conference on Data Engineering (ICDE 2003), 2003, pp. 505–513.
[34] R. Hull, Managing semantic heterogeneity in databases: A theoretical perspective, in: Proceedings of the 16th ACM SIGACT SIGMOD SIGART Symposium on Principles of Database Systems (PODS'97), 1997, pp. 51–61.
[35] H. Katsuno, A.O. Mendelzon, Propositional knowledge base revision and minimal change, Artificial Intelligence 52 (1991) 263–294.
[36] T. Kirk, A.Y. Levy, Y. Sagiv, D. Srivastava, The Information Manifold, in: Proceedings of the AAAI 1995 Spring Symp. on Information Gathering from Heterogeneous, Distributed Environments, 1995, pp. 85–91.
[37] D. Lembo, M. Lenzerini, R. Rosati, Source inconsistency and incompleteness in data integration, in: Proceedings of the 9th International Workshop on Knowledge Representation meets Databases (KRDB 2002), 2002;
CEUR Electronic Workshop Proceedings, http://ceur-ws.org/Vol-54/.
[38] M. Lenzerini, Data integration: A theoretical perspective, in: Proceedings of the 21st ACM SIGACT SIGMOD SIGART Symposium on Principles of Database Systems (PODS 2002), 2002, pp. 233–246.
[39] H.J. Levesque, G. Lakemeyer, The Logic of Knowledge Bases, MIT Press, 2001.

[40] A.Y. Levy, Logic-based techniques in data integration, in: J. Minker (Ed.), Logic Based Artificial Intelligence, Kluwer Academic Publisher, 2000.
[41] J. Lin, A.O. Mendelzon, Merging databases under constraints, Internat. J. Cooperat. Inform. Syst. 7 (1) (1998) 55–76.
[42] I. Manolescu, D. Florescu, D. Kossmann, Answering XML queries on heterogeneous data sources, in: Proceedings of the 27th International Conference on Very Large Data Bases (VLDB 2001), 2001, pp. 241–250.
[43] X. Qian, Query folding, in: Proceedings of the 12th IEEE International Conference on Data Engineering (ICDE'96), 1996, pp. 48–55.
[44] J.D. Ullman, Information integration using logical views, Theoret. Comput. Sci. 239 (2) (2000) 189–210.
[45] M. Winslett, Updating Logical Databases, Cambridge University Press, 1990.